An implicit trust region algorithm for constrained optimization

J. Frederic Bonnans, Geneviève Launay

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.: (1) 39 63 55 11

Rapports de Recherche

1992

25ème anniversaire

N° 1780

Programme 5
Traitement du Signal,
Automatique et Productique

AN IMPLICIT TRUST REGION ALGORITHM FOR CONSTRAINED OPTIMIZATION

Joseph Frédéric BONNANS
Geneviève LAUNAY

Octobre 1992

RR-1780

An implicit trust region algorithm for constrained optimization

Algorithmes de région de confiance implicite pour l'optimisation avec contraintes

J. F. Bonnans – G. Launay[1]

Abstract

In this paper we study the convergence of sequential quadratic programming algorithms for the nonlinear programming problems. Assuming only that the direction is a stationary point of the current quadratic program we study the local convergence properties without strict complementarity. We obtain some global and superlinearly convergent algorithm. As a particular case we formulate an extension of Newton's method that is quadratically convergent to a point satisfying a strong sufficient second order condition.

Résumé

Dans cet article nous étudions la convergence d'algorithmes qui traitent les problèmes de programmation non linéaire par programmation quadratique successive. Notre seule hypothèse étant que la direction est un point stationnaire du problème quadratique, nous étudions des propriétés de convergence locale sans supposer la stricte complémentarité. Nous obtenons un algorithme dont la convergence est globale et superlinéaire. En particulier, nous énonçons une extension de la méthode de Newton dont la convergence est quadratique lorsque le point limite vérifie une condition forte du second ordre.

Key words : Nonlinear programming, Newton's method, quasi-Newton algorithms, exact penalization, trust region.

AMS (MOS) **subject classification** : 90C30, 49M37, 65K05.

[1]INRIA, Domaine de Voluceau, 78153 Rocquencourt, France.

1 Introduction

1.1 The family of Newton type algorithms

In this paper we present a new algorithm for solving the standard nonlinear programming problem

$$\min f(x)\ ;\ g(x) \ll 0, \tag{P}$$

with f, g smooth mapping from $\mathbb{R}^n$ to $\mathbb{R}$ and $\mathbb{R}^p$, and, given a partition (I, J) of $\{1, \ldots, p\}$, by $z \ll 0$ we mean $z_i \leq 0$, $i \in I$, $z_j = 0$, $j \in J$. Occasionally for $K \subset I$ we will denote

$$z \overset{K}{\ll} 0 \Leftrightarrow \begin{cases} z_i \leq 0, & i \in K, \\ z_j = 0, & j \in J. \end{cases}$$

To (P) is associated the first-order optimality system

$$\begin{cases} \nabla f(x) + g'(x)^t \lambda = 0, \\ \\ g(x) \ll 0,\ \lambda_I \geq 0,\ \lambda^t g(x) = 0. \end{cases} \tag{1.1}$$

If (x, λ) satisfies (1.1), then we say that λ is a multiplier associated to x. By extension we say that x is solution of (1.1) if there exists λ such that (x, λ) satisfies (1.1).

We define the quadratic problem

$$\min_d \nabla f(x)^t d + \frac{1}{2} d^t M d\ ;\ g(x) + g'(x) d \ll 0, \tag{$Q(x,M)$}$$

to which is associated the optimality system

$$\begin{cases} \nabla f(x) + Md + g'(x)^t \mu = 0, \\ \\ g(x) + g'(x)d \ll 0,\ \mu_I \geq 0,\ \mu^t(g(x) + g'(x)d) = 0. \end{cases} \tag{1.2}$$

Denote by $L(x, \lambda) := f(x) + \lambda^t g(x)$ the Lagrangian associated to (P). It has been observed by Wilson [19] that, when no inequality is present, the computation of the Newton step in (1.1) amounts to solving d solution of $Q(x, M)$ with $M = \nabla_x^2 L(x, \lambda)$, and that this allows a natural generalization for problems with inequality constraints. In order to deal with the case when second derivatives are not available, a larger class has been defined in the following way :

Algorithm 0 (Newton type algorithms).

0) Choose $x^0 \in \mathbb{R}^n$, M^0 a $n \times n$ symmetric matrix ; $k \leftarrow 0$.
1) Compute (d^k, μ^k) solution of the optimality system of $Q(x^k, M^k)$.
2) Linesearch : choose ρ_k in $[0, 1]$.
3) $x^{k+1} \leftarrow x^k + \rho_k d^k$
   Choose M^{k+1}.
   $k \leftarrow k + 1$, go to 1.

1.2 Some known results on local study

Let $\bar{x}$ be a local solution of (P) to which is associated a unique Lagrange multiplier $\bar{\lambda}$. The local study analysis typically assumes that (x^0, M^0) is close to $(\bar{x}, \nabla^2 L(\bar{x}, \bar{\lambda}))$ and that $\rho_k = 1$. The question is to determine if convergence occurs, and at which rate. It happens that in this case d^k should not, in general, be taken as the global minimum of $Q(x^k, M^k)$.

Indeed, let us consider the simple example

$$\min_x \ell n(1+x) \; ; \; -x \leq 0, \; x \leq 10.$$

This problem has a unique solution $\bar{x} = 0$ associated to the unique multiplier $\bar{\lambda} = (1, 0)^t$ and the strongest regularity hypothesis and sufficient second order condition (see (1.8) and (4.9) below) are satisfied by $(\bar{x}, \bar{\lambda})$. Now let us start Newton's method at the solution : we get the quadratic problem

$$\min_d d - d^2/2 \; ; \; 0 \leq d \leq 10,$$

whose unique solution is $d = 10$, the worst possible displacement ! As the Newton step is obtained by linearizing the data is it clear that the quadratic program is meaningful only if the displacement is not too large. Indeed in our example the "good" displacement $d = 0$ is a local solution of the quadratic program.

Of course if $M^k \geq 0$, which is the case for some quasi-Newton algorithms based on positive definite updates, and also for Newton's method when (P) is convex (i.e. has convex cost and inequality constraints, and linear equality constraints) then $Q(x^k, M^k)$ is itself convex, and local and global minima coincide.

We now quote some recent results about the speed of convergence of Newton type algorithms. For this purpose, we need to define

the set of active inequality constraints:

$$I(x) := \{i \in I \; ; \; g_i(x) = 0\},$$

the set of active constraints $I(x) \cup J$,

the extended critical cone:

$$C(x) := \{d \in \mathbb{R}^n \; ; \; g'(x)d \overset{I(\bar{x})}{\ll} 0 \; ; \; g_i'(x)d = 0 \text{ if } \bar{\lambda}_i > 0, \; i \in I\}. \tag{1.3}$$

Note that when $x = \bar{x}$ we recover the usual critical cone, or cone of critical directions :

$$C(\bar{x}) := \{d \in \mathbb{R}^n \; ; \; g'(\bar{x})d \overset{I(\bar{x})}{\ll} 0 \; ; \; g_i'(\bar{x})d = 0 \text{ if } \bar{\lambda}_i > 0, \; i \in I\}, \tag{1.4}$$

the (standard) second-order sufficient condition:

$$d^t \nabla_x^2 L(\bar{x}, \bar{\lambda}) d > 0, \text{ for all } d \in C(\bar{x}) - \{0\}, \tag{1.5}$$

and the orthogonal projection onto $C(x^k)$, denoted by P^k.

Note that usually the critical cone is defined as

$$C(\bar{x}) := \{d \in \mathbb{R}^n \ ; \ \nabla f(\bar{x})^t d \leq 0 \ ; \ g'(\bar{x})d \overset{I(\bar{x})}{\ll} 0\}.$$

Both definitions coincide (as is easy to check using (1.2)) because we assume the existence of a Lagrange multiplier. We now quote two results of Bonnans [7] :

Theorem 1.1 *Let $\bar{x}$ be a local solution of (1.1) such that the gradients of active constraints are linearly independent, $\bar{\lambda}$ be the unique multiplier associated to $\bar{x}$ and the second-order sufficient condition holds. Then if (x^k, μ^k) computed by Algorithm 0 converge to $(\bar{x}, \bar{\lambda})$, then $\{x^k\}$ converges superlinearly iff*

$$P^k[(\nabla_x^2 L(\bar{x}, \bar{\lambda}) - M^k)d^k] = o(d^k).$$

Theorem 1.2 *Assume that $\bar{x}$ is a local solution of (1.1), $\bar{\lambda}$ is the unique Lagrange multiplier associated to $\bar{x}$, and the second-order sufficiency condition holds. Then there exists $\varepsilon > 0$ such that if $\|x^0 - \bar{x}\| + \|\lambda^0 - \bar{\lambda}\| < \varepsilon$, and (x^{k+1}, λ^{k+1}) is chosen so that $\|x^{k+1} - x^k\| + \|\lambda^{k+1} - \lambda^k\| < 2\varepsilon$, then Algorithm 0 with $M^k = \nabla_x^2 L(x^k, \lambda^k)$ and $\rho_k = 1$ i.e., Newton's method, is well defined and converges at a quadratic rate to $(\bar{x}, \bar{\lambda})$.*

We note that the existence of a unique multiplier is a qualification hypothesis slightly weaker than the linear independance of gradients of active constraints (see [10]). Note also that if the following strict complementarity hypothesis holds:

$$\bar{\lambda}_i > 0 \text{ for all } i \text{ in } I(\bar{x}),$$

then, for (x^k, M^k) close to $(\bar{x}, \nabla^2 L(\bar{x}, \bar{\lambda}))$, λ^k is close to $\bar{\lambda}$, hence if $i \in I(\bar{x})$, the corresponding inequality in $Q(x^k, M^k)$ is active and everything goes as if we were analyzing the problem

$$\min f(x) \ ; \ g_i(x) = 0, \ i \in I(\bar{x}) \cup J.$$

Then Thm 1.1 reduces to a result of Boggs, Tolle and Wang [4], whereas Thm 1.2 reduces to the application of the general result on quadratic convergence of Newton's method for a system of equations. The novelty in the theorems above lies in the fact that no strict complementarity hypothesis holds and only the standard (weak) sufficient condition is assumed.

1.3 Some known results on globalization

The local results that we just presented insure a superlinear or quadratic convergence, provided that the data at the starting point are sufficiently close to the optimum. When these hypotheses are not satisfied the algorithm has to be modified, for different reasons :

(i) It may happen that the optimality system of $Q(x^k, M^k)$ has no local solution ; a possible remedy is to solve a modified quadratic program.

(ii) The point $x^k + d^k$ may be farther from any local solution than x^k. For this reason it is safe to introduce a linesearch on some potential function ; the most popular potential is the Pschenichny-Han exact penalty function (see Han [12], Pschenichny and Danilin [17]) :

$$\theta_r(x) := f(x) + r\|g(x)^\sharp\|$$

with $r > 0$ (the penalty parameter) and $.^\sharp$ defined as follows :

$$z_i^\sharp = \begin{cases} z_i^+ & if \quad i \in I, \\ z_i & if \quad i \in J. \end{cases}$$

Here $\|.\|$ stands for an arbitrary norm in $\mathbb{R}^p$, although we note that most often the ℓ^1 norm is chosen for practical reasons. The dual norm $\|.\|_*$ is defined as

$$\|\mu\|_* := \max\{z^t\mu \ ; \ \|z\| \leq 1\}.$$

Usually r is chosen so that $r > \|\mu^k\|_*$, where μ^k is the multiplier associated to d^k.

However this potential suffers from the Maratos effect (Maratos [14], Mayne and Polak [15]) : even when x is close to $\bar{x}$ and $x + d - \bar{x} = O(x - \bar{x})^2$, and r close to $\|\bar{\lambda}\|_*$, it may happen that $\theta_r(x + d) > \theta_r(x)$, and in the context of composite optimization it has been shown that this may occur an infinite number of time, (see Yuan [20]).

Various remedies have been proposed, the first of them being to make an additional restauration step (Mayne and Polak [15], Gabay [11]), i.e. denoting $\|.\|$ an arbitrary norm in $\mathbb{R}^n$, different from the one in $\mathbb{R}^p$, to compute v^k solution of

$$\min_v \|v\| \ ; \ \begin{cases} g(x^k + d^k) + g'(x^k)v \ll 0, \\ g_i(x^k + d^k) + g'(x^k)v = 0, \ i \in I_k^*, \end{cases}$$

where I_k^* is some prediction of the set of active constraints, obtained as a byproduct of the computation of d^k and to perform a linesearch along the arc

$$\rho \to x^k + \rho d^k + \rho^2 v^k.$$

Other possible remedies are to modify the potential, specifically to use a nondifferentiable augmented Lagrangian [6], and to compare the value of $\theta_r(x^{k+1})$ to the value of θ_r not only at x^k, but also at x^{k-1}, x^{k-2},... (see Panier and Tits [16], Bonnans, Panier, Tits, Zhou [8]).

To the knowledge of the authors, all published papers concerning global convergence (i.e. convergence of at least a subsequence of $\{(x^k, \mu^k)\}$ towards $(\bar{x}, \bar{\lambda})$ satisfying (1.1)) assume that at all iterations $M^k > 0$, as well as its inverse at least in the space spanned by the directions tangent to the linearized constraints, hence do not take in account Newton's method for nonconvex optimization. Also all local studies about Maratos effect assume that the strict complementarity hypothesis holds.

1.4 Our contribution

We present in this paper an algorithm that has global and local properties under weak hypotheses on the sequence $\{M^k\}$ of approximations of the Hessian of the Lagrangian. At step k of the algorithm a parameter $\alpha_k \geq 0$ is set and a direction d^k is computed as a stationary point (if any) of the quadratic problem

$$\min_d \nabla f(x^k)^t d + \frac{1}{2} d^t M^k d + \frac{\alpha_k}{2}\|d\|_2^2 \ ; \ g(x^k) + g'(x^k)d \ll 0 \qquad Q_{\alpha_k}(x^k, M^k)$$

where $\|.\|_2$ is the Euclidean norm.

We note for future reference that the first order optimality system of $Q_{\alpha_k}(x^k, M^k)$ is, denoting by μ^k the Lagrange multiplier :

$$\begin{cases} \nabla f(x^k) + M^k d^k + \alpha_k d^k + g'(x^k)^t \mu^k = 0, \\ g(x^k) + g'(x^k)d^k \ll 0 \; ; \; \mu_I^k \geq 0 \; ; \; (\mu^k)^t(g(x^k) + g'(x^k)d^k) = 0. \end{cases} \tag{1.6}$$

The parameter $r = r_k$ of the exact penalty function $\theta_r(x)$ is adapted at each iteration in order to allow a linesearch ; however, null-steps may happen and in this case α_k is increased. We prove that $\{r_k\}$ and $\{\alpha_k\}$ are bounded and that any limit-point of $\{x^k\}$ satisfies (1.1). Our hypotheses are as follows. First we assume

$$\{M^k\}, \{x^k\} \text{ and } \{d^k\} \text{ are bounded.} \tag{1.7}$$

Note that if upper and lower bounds on x are present, then $\{x^k\}$ and $\{d^k\}$ are necessarily bounded. Second, we assume that

$$\text{the linearized constraints } g(x) + g'(x)d \ll 0 \text{ are qualified,} \tag{1.8}$$

which means that for any (x, d) such that $g(x) + g'(x)d \ll 0$, the gradients of active constraints of this system are linearly independent.

We show also how to adapt some anti-Maratos effects in order to formulate an algorithm that has a superlinear convergence.

If second-order derivatives are available we show how to formulate a globally convergent algorithm that reduces locally to Newton's method, and this seems to be the first globally convergent extension of Newton's idea for constrained optimization.

We call our method an implicit trust region algorithm, making reference to the standard trust region idea in which the subproblem to be solved would be

$$\min_d \nabla f(x^k)^t d + \frac{1}{2} d^t M^k d \; ; \; g(x^k) + g'(x^k)d \ll 0 \; ; \; \frac{1}{2}\|d\|_2^2 \leq \delta_k.$$

The trust region algorithm has two drawbacks. First the computation of its solution necessitates the resolution of several subproblems of type $Q_\alpha(x^k, M^k)$ where now α is interpreted as the multiplier associated to the constraint on the size of d^k. Second, if δ_k is too small, then the constraints may be incompatible. To deal with the last difficulty several authors suggested different relaxation of the constraints (see Fletcher [10]), and these relaxations may indeed be useful if the linearized constraint $g(x) + g'(x)d \ll 0$ may be incompatible at some points. But we have in mind a situation where this does not occur, and it seems to us more natural to formulate the algorithm using nothing else than the subproblem $Q_{\alpha_k}(x^k, M^k)$ that anyway appears in the implementation of the trust region algorithm.

It may seem surprising that the algorithm includes an implicit trust region idea as well as a linesearch; this is due to the presence of constraints. For x fixed, when $\alpha \to \infty$, d solution of $Q_\alpha(x, M)$ converges to $\pi(x)$ solution of

$$\min_d \|d\|_2 \; ; \; g(x) + g'(x)d \ll 0,$$

and (if $\pi(x)$ is nonzero) it may happen that $f(x + \pi(x)) > f(x)$ and $\|g(x + \pi(x))^\sharp\| > \|g(x)^\sharp\|$; in this case the step $\rho_k = 1$ cannot be accepted whenever α_k is large enough.

2 A globally convergent algorithm with fixed penalty parameter

In this section we will present some properties of the exact penalty function that allow to design a linesearch extending the one due to Armijo [1] for unconstrained minimization. The ideas that we present here are classical (see [12]) but we generalize the previous results by assuming that the norm involved in θ_r satisfies only the following property :

$$z \longrightarrow \|z^\sharp\| \quad \text{is a convex mapping} \tag{2.1}$$

We define the directional derivative of θ_r at x in direction d as $\theta_r'(x,d)$. This is well defined, even if (2.1) does not hold, because $\rho \mapsto g(x+\rho d)^\sharp$ has a directional derivative $w(x,d)$ (that can easily be computed explicitely) and $z \longrightarrow \|z\|$ is convex and Lipschitz, hence

$$\begin{aligned}\theta_r(x+\rho d) &= f(x) + \rho f'(x)d + r\|g(x)^\sharp + \rho w(x,d)\| + o(\rho) \\ &= \theta_r(x) + \rho[f'(x)d + r\mu^t w(x,d)] + o(\rho)\end{aligned}$$

where μ is some element of the subdifferential of $\|.\|$ at $g(x)^\sharp$.

We define the "linearized" (at point x^k) exact penalty function as follows :

$$\theta^k(d) = f(x^k) + f'(x^k)d + r_k\|(g(x^k) + g'(x^k)d)^\sharp\|.$$

For any d feasible for $Q_{\alpha_k}(x^k, M^k)$, we note that the decrease of the linearized exact penalty function when step $\rho_k = 1$ is accepted is equal to $\Delta_{r_k}(x^k, d)$, where

$$\Delta_r(x,d) := r\|g(x)^\sharp\| - f'(x)d.$$

We say that $\Delta_r(x,d)$ is feasible if

$$\Delta_r(x,d) \geq \|d\|^3. \tag{2.2}$$

By Δ_k we denote $\Delta_{r_k}(x^k, d^k)$.

Lemma 2.1 *Let d be a stationary point of $Q_\alpha(x,M)$ and μ the associated Lagrange multiplier. Then :*

(i) *If (2.1) holds, then*

$$\theta_r'(x,d) \leq -\Delta_r(x,d). \tag{2.3}$$

(ii) *The following relations hold :*

$$\Delta_r(x,d) \geq (r - \|\mu\|_*)\|g(x)^\sharp\| + \alpha\|d\|_2^2 + d^t M d + \mu^t(g(x)^\sharp - g(x)) \tag{2.4}$$

$$\Delta_r(x,d) \geq (r - \|\mu\|_*)\|g(x)^\sharp\| + \alpha\|d\|_2^2 + d^t M d. \tag{2.5}$$

Proof

(i) From (2.1) we deduce that

$$\theta'_r(x,d) = f'(x)d + r\eta^t g'(x)d$$

where η is some subgradient of $\|.^\sharp\|$ at $g(x)$, i.e.

$$\|z^\sharp\| \geq \|g(x)^\sharp\| + \eta^t(z - g(x)), \quad \forall z \in \mathbb{R}^p.$$

Choosing $z = g(x) + g'(x)d$, and noting that $z^\sharp = 0$, we deduce that $\eta^t g'(x)d \leq -\|g(x)^\sharp\|$, from which (2.3) follows.

(ii) From (1.6) we deduce

$$0 = f'(x)d + d^t M d + \alpha\|d\|_2^2 + \mu^t g'(x)d.$$

From the complementarity condition we get that $\mu^t g'(x)d = -\mu^t g(x)$, hence

$$-f'(x)d = d^t M d + \alpha\|d\|_2^2 - \mu^t g(x),$$

and so

$$\begin{aligned}
\Delta_r(x,d) &= \alpha\|d\|_2^2 + d^t M d + r\|g(x)^\sharp\| - \mu^t g(x), \\
&= \alpha\|d\|_2^2 + d^t M d + r\|g(x)^\sharp\| - \mu^t g(x)^\sharp + \mu^t(g(x)^\sharp - g(x)) \\
&\geq \alpha\|d\|_2^2 + d^t M d + (r - \|\mu\|_*)\|g(x)^\sharp\| + \mu^t(g(x)^\sharp - g(x)).
\end{aligned}$$

Thus (2.4) is proved. Now, as $\mu_I \geq 0$, we get from the definition of $g(x)^\sharp$ that $\mu^t(g(x)^\sharp - g(x)) \geq 0$, and so (2.5) holds. □

Let x^k be the current point of the algorithm and d^k a stationary point of $Q_{\alpha_k}(x^k, M^k)$. From (2.5) it follows that, at least if $r_k > \|\mu^k\|_*$ and α_k is large enough, then Δ_k is feasible (note that for α_k sufficiently large, $\|d^k\| \approx \|\pi(x^k)\|$, hence (2.2) is satisfied).

From (2.3) it follows that d^k is a descent direction of θ_{r_k} if $\Delta_k > 0$. This allows to define a linesearch in the following way :

Linesearch rule LS1: Parameters $\gamma \in (0, 1/2), \beta \in (0,1)$.

If Δ_k is feasible then compute $\rho_k = (\beta)^\ell$, with ℓ smallest integer such that

$$\begin{aligned}
&\theta_{r_k}(x^k + (\beta)^\ell d^k) \leq \theta_{r_k}(x^k) - (\beta)^\ell \gamma \Delta_k \\
&x^{k+1} \leftarrow x^k + \rho_k d^k.
\end{aligned} \tag{2.6}$$

We note that (2.3) and the relation $\gamma < 1/2$ imply that (2.6) is satisfied for ℓ large enough. Hence the linesearch is well defined. In order to analyse the global properties associated with this linesearch we deal in this section with the simple case when r_k is equal to some constant r.

We can now formulate a conceptual algorithm :

Algorithm 1

0) Data : $\alpha_0 \geq 0$, $M^0 n \times n$ symmetric matrix, $x^0 \in \mathbb{R}^n$; $k \to 0$

1) Computation of (d^k, μ^k) satisfying the optimality system of $Q_{\alpha_k}(x^k, M^k)$

2) If Δ_k is not feasible (i.e. (2.2) not satisfied for Δ_k), stop.
3) Perform the linesearch LS1
4) Choose α_{k+1} and M^{k+1} ;
   $k \to k+1$,
   go to 1).

Theorem 2.1 *Assume that (1.7) and (1.8) hold. Let x^k be computed by Algorithm 1 in which Δ_k is assumed to be feasible at each step. Assume that (α_k, M^k, d^k) are bounded, $r_k = r > 0$. Then $d^k \to 0$ and the set of limit points of (x^k, μ^k) is a connex subset of the set of solutions of the first-order optimality system (1.1).*

Proof We prove that $d^k \to 0$. We note that $\theta_r(x^k)$ decreases hence converges, so that by (2.6) $\rho_k\Delta_k \to 0$. Assume that for some subsequence k' we have $(x^{k'}, \alpha_{k'}, M^{k'}, d^{k'}) \to (\hat{x}, \hat{\alpha}, \hat{M}, \hat{d})$. with $\hat{d} \neq 0$. We observe that $\Delta_{k'} \to \hat{\Delta} := \Delta_r(\hat{x}, \hat{d}) > 0$ by (2.2) and that $\hat{d}$ satisfies the first-order optimality system of $Q_{\hat{\alpha}}(\hat{x}, \hat{M})$; hence $\theta_r'(\hat{x}, \hat{d}) \leq -\hat{\Delta}$ by (2.3), which implies for ρ small enough

$$\begin{aligned} \theta_r(\hat{x} + \rho\hat{d}) &\leq \theta_r(\hat{x}) - \rho\hat{\Delta} + o(\rho) \\ &\leq \theta_r(\hat{x}) - \frac{2\rho}{3}\hat{\Delta}, \end{aligned}$$

hence for k' large enough by continuity (as $\Delta_{k'} \to \hat{\Delta} > 0$):

$$\theta_r(x^{k'} + \rho d^{k'}) \leq \theta_r(x^{k'}) - \frac{\rho}{2}\Delta_r(x^{k'}, d^{k'})$$

which proves that $\rho_{k'}$ cannot converge to 0, hence we get $\hat{\Delta} = \lim \Delta_{k'} = 0$, from $\rho_k\Delta_k \to 0$, contradicting $\hat{\Delta} > 0$ obtained from our assumption $\hat{d} \neq 0$.

Now as $d^k \to 0$ for any converging subsequence of $(x^k, \alpha_k, M^k, d^k)$ we can pass to the limit in (1.6), deducing the boundedness of $\{\mu^k\}$ from (1.7) and (1.8), and so that any limit point of ((x^k, μ^k) is solution of (1.1). Now as $d^k \to 0$, the set of limit points of $\{x^k\}$ is connex; by (1.8) the Lagrange multiplier of (1.1) (whenever it exists) must depend continuously on x; the conclusion follows. □

In the next section we relax the restrictive hypothesis on r^k and on the a priori feasibility of Δ_k.

3 A general globally convergent algorithm

This section is devoted to the statement and analysis of a globally convergent algorithm, i.e. an algorithm computing a sequence $\{x^k, \mu^k\}$ such that any of its limit-points satisfies the first-order optimality conditions (1.1). In this algorithm we have to update the two parameters r_k and α_k.

For r_k the idea is the following : take $r_k = r_{k-1}$ whenever it is possible, i.e. if $\Delta_{r_{k-1}}(x^k, d^k)$ is feasible and $\rho_k = 1$ is accepted by the linesearch; otherwise choose r_k satisfying $r_k >$

$\|\mu^k\|_*$; in order to make the sequence r_k constant after a finite number of step we choose $r_k = \max(r_{k-1}, int(\|\mu^k\|_* + 2))$. Finally the update rule for r_k is as follows :

$$r_k = \begin{cases} r_{k-1} \text{ if } \Delta_{r_{k-1}}(x^k, d^k) \text{ is feasible and } \theta_{r_{k-1}}(x^k + d^k) \leq \theta_{r_{k-1}}(x^k) - \gamma \Delta_{r_{k-1}}(x^k, d^k), \\ \\ \max(r_{k-1}, int(\|\mu^k\|_* + 2)) \text{ if not.} \end{cases} \tag{3.1}$$

For α_k the idea is the following : if Δ_k is not feasible or ρ_k is close to 0, then choose $\alpha_{k+1} > \alpha_k + \varepsilon_1$, with $\varepsilon_1 > 0$ (because of Lemma 2.1 this will eventually allow to get feasibility of Δ_k). On the other hand, if Δ_k is feasible and $\rho_k = 1$ then α_{k+1} will be taken smaller than α_k.

Finally we mention the possibility of null-steps, i.e. when Δ_k is not feasible then x^{k+1} is taken equal to x^k (or equivalently $\rho_k = 0$). We now state the algorithm:

Algorithm 2

0) Data : $\alpha_0 \geq 0$, M^0 $n \times n$ symmetric matrix, $x^0 \in \mathbb{R}^n$. Parameters $0 < \varepsilon_1 < \varepsilon_2$, $0 < \varepsilon_3 < 1$; $k \leftarrow 0$.
1) Computation of (d^k, μ^k), satisfying the optimality system of $Q_{\alpha_k}(x^k, M^k)$.
2) If $k = 0$, set $r_{-1} \leftarrow \|\mu^0\|_* + 1$.
3) Choice of r_k using the rule (3.1).
4) If Δ_k is not feasible (null step) :
   $\rho_k \leftarrow 0$,
   $x^{k+1} \leftarrow x^k$,
   go to 6).
5) If Δ_k is feasible : perform the linesearch LS1.
6) Update of α_k :
   If $\rho_k = 1$, choose $\alpha_{k+1} \leq \alpha_k/2$.
   If $\rho_k \in (\varepsilon_3, 1)$, choose $\alpha_{k+1} \leq \alpha_k + \varepsilon_2$.
   If $\rho_k \leq \varepsilon_3$ choose $\alpha_{k+1} \in [\alpha_k + \varepsilon_1, \alpha_k + \varepsilon_2]$.
   Choose M^{k+1}.
7) $k \leftarrow k + 1$,
   go to 1). □

Remark 3.1 *We observe that $\{r_k\}$ increases, and $\{r_k\}$ is bounded iff $r_k = r$ for $k \geq k_0$.*

Theorem 3.1 *Let x^k be computed by Algorithm 2. We assume that (1.7) and (1.8) hold. Then :*

(i) *The sequences $\{r_k\}$, $\{\alpha_k\}$ and $\{\mu^k\}$ are bounded,*

(ii) *The set of limit-points of $\{x^k\}$ is connex, and to each limit point is associated a Lagrange multiplier.*

We give a proof that makes use of some lemmas below.

Proof
a) We prove that $\{r_k\}$ is bounded. If not, then there exists a subsequence k' with $r_{k'} > r_{k'-1}$, and by (3.1) $\|\mu^{k'}\|_* \to \infty$. This, and (1.6) - (1.8) imply that $\alpha_{k'}\|d^{k'}\| \to \infty$. Now Lemma 3.1 proves that $\|g(x^k)^\sharp\| \to 0$ and Lemma 3.2 ensures that for k' large enough, $r_{k'} = r_{k'+1}$, in contradiction with the definition of $\{k'\}$.

b) We prove that $\{\alpha_k\}$ is bounded. As $\{r_k\}$ is bounded, we know from Remark 3.1 that r_k is constant, say equal to r for $k \geq k_0$. Lemma 3.3 says that Δ_k is feasible if $\alpha_k \geq \hat{\alpha}$ and $k \geq k_0$. From Algorithm 2 it follows that $\alpha_{k+1} \leq \alpha_k + \varepsilon_2$ for all k and by Lemma 3.3 if $\alpha_k \geq \hat{\alpha}$ and $k \geq k_0$, then $\rho_k = 1$ and $\alpha_{k+1} \leq \alpha_k/2$: hence $\alpha_{k+1} \leq \max(\hat{\alpha}, \alpha_{k_0}/2) + \varepsilon_2$ whenever $k \geq k_0$.

c) We now prove (ii) : this follows applying Theorem 2.1, taking in account that after at most $\hat{\alpha}/\varepsilon_1$ successive null steps one has $\alpha_k \geq \hat{\alpha}$. We now may apply Theorem 2.1 to the renumbered sequence obtained by deleting the null steps, and this is an infinite sequence, whose limit-points are the same as those of the original sequence. $\square$

We now state and prove the three lemmas used in the proof of Thm 3.1.

Lemma 3.1 *Let $\{x^k\}$ be computed by Algorithm 2. Under hypotheses(1.7) and (1.8), if $r_k \nearrow \infty$ then $\|g(x^k)^\sharp\| \to 0$.*

Proof
a) Let us verify that $\|g(x^k)^\sharp\|$ converges. Let $m := inf\{f(x^k), k \in \mathbb{N}\}$. Note that $m > -\infty$ as $\{x^k\}$ is bounded. Then, as $\{r_k\}$ increases $\theta_{r_k}(x^{k+1}) \leq \theta_{r_k}(x^k)$ and so we deduce

$$\|g(x^{k+1})^\sharp\| + \frac{f(x^{k+1}) - m}{r_k} \leq \|g(x^k)^\sharp\| + \frac{f(x^k) - m}{r_k} \leq \|g(x^k)^\sharp\| + \frac{f(x^k) - m}{r_{k-1}},$$

hence $\{\|g(x^k)^\sharp\| + (f(x^k) - m)/r_{k-1}\}$ is a decreasing sequence, and so converges since it is bounded. As $r_k \nearrow \infty$ and $\{f(x^k)\}$ is bounded since $\{x^k\}$ is bounded, it follows that $\|g(x^k)^\sharp\|$ converges.

b) It suffices now to get a contradiction when assuming that $\lim\|g(x^k)^\sharp\|$ is positive. Let us note that by (1.6) - (1.8), if $\{\alpha_k\}$ is bounded, so is $\{\mu^k\}$ hence $\{r_k\}$ cannot go to ∞. Hence we may extract a subsequence k' such that $\alpha_{k'} \to \infty$ and $x^{k'} \to \bar{x}$.It is easily checked that $d^{k'} \to \bar{d} := \pi(\bar{x})$. Now since $\|.^\sharp\|$ is a Lipschitz mapping:

$$\begin{aligned}
\theta_{r_{k'}}(x^{k'}) - \theta_{r_{k'}}(x^{k'} + \rho d^{k'}) &= r_{k'}\|g(x^{k'})^\sharp\| - r_{k'}\|g(x^{k'} + \rho d^{k'})^\sharp\| + O(1) \\
&= r_{k'}\|g(\bar{x})^\sharp\| - r_{k'}\|g(\bar{x} + \rho\bar{d})^\sharp\| + o(r_{k'}),
\end{aligned}$$

with $o(r_{k'})/r_{k'} \to 0$ uniformly on $\rho \in [0,1]$.

As $g(x^{k'}) + g'(x^{k'})d^k \ll 0$ and (1.7) holds, it follows that

$$\|g(\bar{x} + \rho\bar{d})^\sharp\| \leq (1-\rho)\|g(\bar{x})^\sharp\| + a_0\rho^2 \text{ for some } a_0 > 0,$$

hence since $\|.^\sharp\|$ is a Lipschitz mapping and (1.7) holds:

$$\begin{aligned}
\theta_{r_{k'}}(x^{k'}) - \theta_{r_{k'}}(x^{k'} + \rho d^{k'}) &\geq \rho r_{k'}\|g(\bar{x})^\sharp\| - r_{k'}a_0\rho^2 + o(r_{k'}), \\
&\geq \rho r_{k'}\|g(x^{k'})^\sharp\| - r_{k'}a_0\rho^2 + o(r_{k'}), \\
&= \rho\Delta_{k'} - r_{k'}a_0\rho^2 + o(r_{k'}).
\end{aligned}$$

We note that $\Delta_{k'}/r_{k'} \to \|g(\bar{x})^\sharp\|$ which is assumed to be positive. Using this we get for some $a_1 > 0$

$$\theta_{r_{k'}}(x^{k'}) - \theta_{r_{k'}}(x^{k'} + \rho d^{k'}) \geq \Delta_{k'}[\rho - a_1\rho^2 + o(1)]$$

and it follows that $\rho_{k'} \geq \hat{\rho}$ for some $\hat{\rho} > 0$. Then this implies that for some $a_2 > 0$

$$\overline{\lim}\|g(x^{k'+1})^\sharp\|/\|g(x^{k'})^\sharp\| \leq 1 - a_2\hat{\rho},$$

in contradiction with our hypothesis. □

Lemma 3.2 *Let x^k be computed by Algorithm 2. Under the hypotheses (1.7) and (1.8), if a subsequence $\{x^{k'}\}$ satisfies $\|g(x^{k'})^\sharp\| \to 0$ and $\alpha_{k'}\|d^{k'}\| \to \infty$, then*

(i) $\|d^{k'}\|_2/\|\pi(x^{k'})\|_2 \to 1$,

(ii) *For k' large enough, $r_{k'} = r_{k'-1}$.*

Proof Denote by

$$q_k(d) := \nabla f(x^k)^t d + \frac{1}{2}d^t M^k d + \frac{1}{2}\alpha_k d^t d$$

the cost function of $Q_{\alpha_k}(x^k, M^k)$. As $\|d^k\|$ is bounded it follows from the unboundedness of $\alpha_{k'}\|d^{k'}\|$ that $\alpha_{k'} \to \infty$. So we see that for $k' \geq k'_0$, $q_{k'}(d)$ is convex, hence $d^{k'}$ is a global solution of $Q_{\alpha_{k'}}(x^{k'}, M^{k'})$. In particular, denoting $\pi^k := \pi(x^k)$, we have

$$q_{k'}(d^{k'}) \leq q_{k'}(\pi^{k'}). \tag{3.2}$$

From the definition of π^k we have $\|\pi^k\|_2 \leq \|d^k\|_2$. On the other hand, dividing (3.2) by $\alpha_{k'}\|d^{k'}\|_2^2$, remembering that $\alpha_{k'} \to \infty$ we obtain $1 \leq \underline{\lim}\|\pi^{k'}\|_2/\|d^{k'}\|_2$ and point (i) follows.

We now prove (ii). We may assume that $r_k \nearrow \infty$, because otherwise r_k is constant for k large enough (see Remark 3.1) and then the conclusion holds trivially. The idea is that the penalization term dominates in the linesearch. Indeed,

$$\begin{aligned} \Delta_{r_{k'-1}}(x^{k'}, d^{k'}) &= r_{k'-1}\|g(x^{k'})^\sharp\| - f'(x^{k'})d^{k'} \\ &= r_{k'-1}\|g(x^{k'})^\sharp\| \left(1 - \frac{1}{r_{k'-1}} \frac{f'(x^{k'})d^{k'}}{\|\pi^{k'}\|_2} . \frac{\|\pi^{k'}\|_2}{\|g(x^{k'})^\sharp\|}\right) \end{aligned}$$

Extracting if necessary a subsequence we get $x^{k'} \to \hat{x}$ feasible and so $\|\pi^{k'}\| \to 0$.

Now for the considered sequence, $f'(x^{k'})d^{k'}/\|\pi^{k'}\|_2$ is bounded (by point (i)) and $\|\pi^{k'}\|_2/\|g(x^{k'})^\sharp\|$ is also bounded as a consequence of the qualification hypothesis (1.8), hence

$$\Delta_{r_{k'-1}}(x^{k'}, d^{k'}) = \|g(x^{k'})^\sharp\|(r_{k'-1} + o(r_{k'-1})). \tag{3.3}$$

Also by (i), $\|d^{k'}\|_2/\|g(x^{k'})^\sharp\|$ is bounded. As $\|g(x^{k'})^\sharp\| \to 0$ it follows that $\Delta_{r_{k'-1}}(x^{k'}, d^{k'})$ is feasible.

On the other hand

$$\theta_{r_{k'-1}}(x^{k'}) - \theta_{r_{k'-1}}(x^{k'} + d^{k'}) = r_{k'-1}(\|g(x^{k'})^\sharp\| - \|g(x^{k'} + d^{k'})^\sharp\|) + f(x^{k'}) - f(x^{k'} + d^{k'})$$

and so from (3.3) :

$$\theta_{r_{k'-1}}(x^{k'}) - \theta_{r_{k'-1}}(x^{k'} + d^{k'}) = \Delta_{r_{k'-1}}(x^{k'}, d^{k'}) - r_{k'-1}\|g(x^{k'} + d^{k'})^\sharp\| + O(d^{k'})^2. \tag{3.4}$$

But $\|.^\sharp\|$ is a Lipschitz mapping, and from (1.6) $(g(x^{k'})+g'(x^{k'})d^{k'})^\sharp = 0$, hence $\|g(x^{k'}+d^{k'})^\sharp\| = O(d^{k'})^2$. Also $d^{k'} = O(g(x^{k'})^\sharp)$ hence , with (3.4) :

$$\theta_{r_{k'-1}}(x^{k'}) - \theta_{r_{k'-1}}(x^{k'} + d^{k'}) = \Delta_{r_{k'-1}}(x^{k'}, d^{k'}) + o(\Delta_{r_{k'-1}}(x^{k'}, d^{k'})).$$

As the rule (3.1) is used in Algorithm 2, the two previous results imply $r_{k'-1} = r_{k'}$ for any $k' \geq k_0'$, in contradiction with the hypothesis $r_{k'} \nearrow \infty$. □

Lemma 3.3 *Let x^k be computed by Algorithm 2. Under hypotheses (1.7) and (1.8), if $\{r_k\}$ is bounded, then there exists $\hat{\alpha} > 0$ and k_0 such that $\rho_k = 1$ whenever $\alpha_k \geq \hat{\alpha}$ and $k \geq k_0$.*

Proof Since r_k is bounded there exists r such that $r_k = r$ for $k \geq k_0$ (cf. Remark 3.1).Using (2.5) we know that

$$\Delta_k \geq (r - \|\mu^k\|_*)\|g(x^k)^\sharp\| + \alpha_k\|d^k\|_2^2 + d^{kt}M^k d^k,$$

and so, as from (1.7) $\{M^k\}$ is bounded, we obtain for some $a_3 > 0$:

$$\Delta_k \geq (r - \|\mu^k\|_*)\|g(x^k)^\sharp\| + (\alpha_k - a_3)\|d^k\|_2^2,$$

for k large enough. If $\rho_k \neq 1$ then $r_k = r > \|\mu^k\|_*$; hence

$$\Delta_k \geq (\alpha_k - a_3)\|d^k\|_2^2. \tag{3.5}$$

As $\{d^k\}$ is bounded, we deduce that for α_k large enough, Δ_k is feasible. Now

$$\theta_r(x^k) - \theta_r(x^k+d^k) = r(\|g(x^k)^\sharp\| - \|g(x^k + d^k)^\sharp\|) + f(x^k) - f(x^k + d^k),$$

so since (1.7) holds and f, g are smooth, we get for some $a_4 > 0$:

$$\theta_r(x^k) - \theta_r(x^k + d^k) \geq \Delta_k - a_4\|d^k\|_2^2,$$

hence using (3.5), for α_k large enough :

$$\theta_r(x^k) - \theta_r(x^k+d^k) \geq \frac{1}{2}\Delta_k.$$

As $\gamma < \frac{1}{2}$, the rule (3.1) ensures that the two previous results imply $\rho_k = 1$, in contradiction with the hypothesis $\rho_k \neq 1$. □

4 A globally and superlinearly convergent algorithm

Let $\bar{x}$ be a local solution of (P) and $\bar{\lambda}$ its associated Lagrange multiplier. We know that Algorithm 2 is not in general superlinearly convergent, even if $x^k \to \bar{x}$ and $M^k \to \nabla_x^2 L(\bar{x},\bar{\lambda})$. This is due to the Maratos effect (Maratos [14], Mayne and Polak [15]). In this section we show how to adapt the idea of a restauration step in order to accept the unit stepsize. We define

$$\begin{aligned} I^* &:= \{i \in I\ ;\ \bar{\lambda}_i > 0\} \cup J. \\ I_k^* &:= \{i \in I\ ;\ \mu_i^k > 0\} \cup J. \end{aligned}$$

We first perform a local analysis in which our hypotheses are as follows :

$$\left\{ \begin{array}{l} \{M^k\},\ \{x^k\},\ \{\alpha_k\},\ \{d^k\} \text{ are given such that} \\ x^k \to \bar{x}, \\ \{M^k\} \text{ and } \{\alpha_k\} \text{ are bounded,} \\ d^k \text{ is stationary point of } Q_{\alpha_k}(x^k, M^k) \\ \text{and } d^k \to 0 \end{array} \right\} \tag{4.1}$$

We define v^k as the solution of:

$$\min_v \|v\|\ ;\ \left\{ \begin{array}{l} g(x^k+d^k) + g'(x^k)v \ll 0, \\ g_i(x^k+d^k) + g_i'(x^k)v = 0 \text{ for any } i \in I_k^*; \end{array} \right. \tag{4.2}$$

where $\|.\|$ is an arbitrary norm in $\mathbb{R}^n$. Under some reasonable assumptions we show in Prop. 4.1. below that the point $x^k+d^k+v^k$ insures a significant decrease of the exact penalty function. We start with a technical lemma.

Lemma 4.1 *Assume that (1.8), (4.1) and (4.2) hold. Then one has for some $a > 0$*

$$I^* \subset I_k^* \textit{ for } k \textit{ large enough,} \tag{4.3}$$

$$\|v^k\| \le a\|d^k\|^2 \tag{4.4}$$

$$g(x^k+d^k+v^k)^\sharp = o(d^k)^2, \tag{4.5}$$

$$g_{I^*}(x^k+d^k+v^k) = o(d^k)^2. \tag{4.6}$$

Proof

a) It follows from (4.1), (1.6) and (1.8) that $\mu^k \to \bar{\lambda}$. So for k large enough, $\{i \in I\ ;\ \bar{\lambda}_i > 0\} \subset \{i \in I\ ;\ \mu_i^k > 0\}$ and thus (4.3) is proved.

b) Since (1.8) holds and by definition of I_k^* :

$$\begin{array}{l} g(x^k) + g'(x^k)d^k \ll 0 \\ g_i(x^k) + g_i'(x^k)d^k = 0,\ i \in I_k^* \end{array}$$

it follows that

$$\begin{aligned} g(x^k+d^k) &\ll O(d^k)^2, \\ g_i(x^k+d^k) &= O(d^k)^2,\ i \in I_k^* \end{aligned}$$

hence using again (1.8) , $v^k = O(d^k)^2$.

c) Expanding $g(x^k + d^k + v^k)$ and using (4.4) we get

$$g(x^k + d^k + v^k) = g(x^k) + g'(x^k)(d^k + v^k) + \frac{1}{2}(d^k)^t g''(x^k)d^k + o(d^k)^2 \tag{4.7}$$

Moreover, since (4.2) implies $(g(x^k + d^k) + g'(x^k)v^k)^\sharp = 0$, expanding $g(x^k + d^k)$ and using $z \to \|z^\sharp\|$ Lipschitz we obtain

$$\|(g(x^k) + g'(x^k)d^k + \frac{1}{2}(d^k)^t g''(x^k)d^k + g'(x^k)v^k)^\sharp\| = o(d^k)^2.$$

Then, as $z \to \|z^\sharp\|$ is Lipschitz, we have (4.5).

d) Since v^k is solution of (4.2), the expansion of $g_{I_k^*}(x^k + d^k)$ yields

$$g_i(x^k) + g_i'(x^k)d^k + \frac{1}{2}(d^k)^t g_i''(x^k)d^k + g_i'(x^k)v^k = o(d^k)^2 \text{ for any } i \in I_k^*.$$

Hence (4.6) follows from (4.3) and (4.7). □

Then we compute x^{k+1} along the path $\rho \to x^k + \rho d^k + \rho^2 v^k$. The first trial point is $x^k + d^k + v^k$ and if it appears to be necessary to test a small value for ρ_k, then the contribution of v^k is small with respect to the one of d^k, and this allows to preserve the descent property on θ_r. Specifically the linesearch is as follows :

Linesearch rule LS2 : Parameters $\gamma \in (0, 1/2)$, $\beta \in (0, 1)$.

Compute v^k solution of (4.2).
If Δ_k is feasible, i.e. (2.2) holds for Δ_k, then compute $\rho_k = (\beta)^\ell$ with ℓ smallest integer such that

$$\begin{aligned} &\theta_{r_k}(x^k + (\beta)^\ell d^k + (\beta)^{2\ell} v^k) \leq \theta_{r_k}(x^k) - (\beta)^\ell \gamma \Delta_k, \\ &x^{k+1} \leftarrow x^k + \rho_k d^k + (\rho_k)^2 d^k. \end{aligned} \tag{4.8}$$

In order to perform a local analysis we are led to assume that $(\bar{x}, \bar{\lambda})$ (local solution (P) and associated multiplier) satisfies the following strong second-order sufficient condition :

$$\text{for any} \quad d \in \ker g'_{I^*}(\bar{x}) - \{0\}, \quad d^t \nabla_x^2 L(\bar{x}, \bar{\lambda})d > 0. \tag{4.9}$$

Recalling (1.8) we see that (4.9) is stronger than the standard sufficient condition (1.5), and that both coincide if the strict complementarity hypothesis holds at $\bar{x}$.

The following proposition insures that, locally and under this strong second-order condition (4.9), the restauration increment v^k allows to accept the step $\rho_k = 1$ (i.e. Δ_k is feasible and the exact penalty function decreases sufficiently). We define

$$\begin{aligned} &d_T^k \text{ orthogonal projection of } d^k \text{ onto } \ker g'_{I^*}(x^k), \\ &d_N^k := d^k - d_T^k, \\ &H := \nabla_x^2 L(\bar{x}, \bar{\lambda}), \end{aligned}$$

and for $z \in \mathbb{R}^p, \tilde{z}$ by

$$\tilde{z}_i = \begin{cases} z_i & \text{if } i \in I^*, \\ z_i^+ & \text{if not.} \end{cases}$$

Proposition 4.1 *Assume $\{M^k\}$, $\{x^k\}$, $\{\alpha_k\}$, $\{r_k\}$, $\{d^k\}$ given such that (1.8), (4.1), (4.2) and (4.9) hold and $r_k = r$ with $r > \|\bar{\lambda}\|_*$. If*

$$(d_T^k)^t M^k d_T^k + \alpha_k \|d_T^k\|^2 \geq (d_T^k)^t H d_T^k + o(d_T^k)^2 \tag{4.10}$$

holds, then LS2 accepts the step $\rho_k = 1$ for k large enough.

Before giving the proof we set some preliminary results:

Lemma 4.2 *For any $n \times n$ symmetric matrix M and for any $\varepsilon > 0$ one has :*

$$(d^k)^t M d^k \;\geq\; (d_T^k)^t M d_T^k - \varepsilon^2 \|d_T^k\|_2^2 - \|M\|(1 + \|M\|/\varepsilon^2)\|d_N^k\|_2^2 \tag{4.11}$$

$$(d_T^k)^t M d_T^k \;\geq\; (d^k)^t M d^k - \varepsilon^2 \|d_T^k\|_2^2 - \|M\|(1 + \|M\|/\varepsilon^2)\|d_N^k\|_2^2 \tag{4.12}$$

Proof Since $d^k = d_T^k + d_N^k$ we get

$$(d^k)^t M d^k = (d_T^k)^t M d_T^k + 2(d_T^k)^t M d_N^k + (d_N^k)^t M d_N^k,$$

hence the two following relations hold:

$$(d^k)^t M d^k \geq (d_T^k)^t M d_T^k - 2\|M\|\|d_T^k\|_2 \|d_N^k\|_2 - \|M\|\|d_N^k\|_2^2, \tag{4.13}$$

$$(d_T^k)^t M d_T^k \geq (d^k)^t M d^k - 2\|M\|\|d_T^k\|_2 \|d_N^k\|_2 - \|M\|\|d_N^k\|_2^2. \tag{4.14}$$

As for all $\varepsilon > 0$, $a > 0$, $b > 0$, one has $2ab = 2(\varepsilon a)(b/\varepsilon) \leq \varepsilon^2 a^2 + b^2/\varepsilon^2$, it comes for $a = \|d_T^k\|_2$ and $b = \|d_N^k\|_2 \|M\|$:

$$-2\|M\|\|d_T^k\|_2 \|d_N^k\|_2 \geq -\varepsilon^2 \|d_T^k\|_2^2 - \|M\|^2 \|d_N^k\|_2^2/\varepsilon^2,$$

which with (4.13) and (4.14) gives the conclusion. □

Lemma 4.3 *Under the hypotheses of Proposition 4.1 , for k large enough Δ_k is feasible and the following holds:*

$$\exists a_5 > 0 \; ; \; \Delta_k \geq a_5 \|d^k\|_2^2 \tag{4.15}$$

and (γ being the constant involved in LS2, i.e. $\gamma \in (0, 1/2)$):

$$\Delta_k \geq \frac{1}{2(1-\gamma)} (d^k)^t H d^k + o(d^k)^2 \tag{4.16}$$

Proof a) Preliminaries :

It was already noticed (cf. Proof of Lemma 4.1.a)) that under our hypotheses $\mu^k \to \bar{\lambda}$. From this result and the hypothesis $r > \|\bar{\lambda}\|_*$ one has for k large enough

$$r - \|\mu^k\|_* \geq (r - \|\bar{\lambda}\|_*)/2,$$

and also it comes for $\zeta := \min\{\bar{\lambda}_i \; ; \; i \in I^* \cap I\}$ (and so $\zeta > 0$) that for k large enough

$$\min\{\mu_i^k \; ; \; i \in I^* \cap I\} > \frac{\zeta}{2}$$

Hence, as $\mu^k \geq 0$ and $g(x^k)^\sharp \geq g(x^k)$,

$$\begin{aligned}(\mu^k)^t(g(x^k)^\sharp - g(x^k)) &\geq \frac{\zeta}{2}\sum_{i\in I^\bullet\cap I}(g_i(x^k)^\sharp - g_i(x^k)) \\ &= \frac{\zeta}{2}\sum_{i\in I^\bullet\cap I}\max(0,-g_i(x^k))\end{aligned}$$

From the definition of $g(x^k)^\sharp$ and $\tilde{g}(x^k)$ we finally get with (2.4) that there exists $\xi > 0$ such that for k large enough :

$$\Delta_k \geq \xi\|\tilde{g}(x^k)\| + \alpha_k\|d^k\|^2 + (d^k)^t M^k d^k.$$

Now from (4.11) of Lemma 4.2 with $M = M^k$ it follows that for all $\varepsilon > 0$:

$$\Delta_k \geq \xi\|\tilde{g}(x^k)\| + \alpha_k\|d^k\|_2^2 + (d_T^k)^t M^k d_T^k - \varepsilon^2\|d_T^k\|_2^2 - \|M^k\|(1+\|M^k\|/\varepsilon^2)\|d_N^k\|_2^2.$$

As $\{M^k\}$ is bounded, (1.8) holds, and d_N^k is solution of

$$\min_d \|d\|_2\ ;\ g_{I^\bullet}(x^k) + g'_{I^\bullet}(x^k)d = 0,$$

we have

$$d_N^k = O(g_{I^\bullet}(x^k)) = O(\tilde{g}(x^k)) \tag{4.17}$$

hence for k large enough, since $\|d^k\|_2^2 = \|d_T^k\|_2^2 + \|d_N^k\|_2^2 \geq \|d_T^k\|_2^2$, we get

$$\Delta_k \geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + (\alpha_k - \varepsilon^2)\|d_T^k\|_2^2 + (d_T^k)^t M^k d_T^k \tag{4.18}$$

b) Proof of (4.15) : Since (1.8) and (4.9) hold, there exists $\delta > 0$ such that for x^k close enough to $\bar{x}$:

$$\text{for any}\quad d \in \ker g'_{I^\bullet}(x^k),\quad d^t H d \geq \delta\|d\|^2. \tag{4.19}$$

From (4.10), (4.18) and (4.19) one has for k large enough :

$$\begin{aligned}\Delta_k &\geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + (d_T^k)^t H d_T^k - \varepsilon^2\|d_T^k\|_2^2 + o(d_T^k)^2, \\ &\geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + (\delta - \varepsilon^2)\|d_T^k\|_2^2 + o(d_T^k)^2.\end{aligned}$$

Hence for k large enough, taking $\varepsilon = \sqrt{\delta/3}$ we get :

$$\Delta_k \geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + \frac{\delta}{2}\|d_T^k\|_2^2.$$

Using (4.17) we deduce (4.15). Hence, as we assume that $d^k \to 0$, it follows that Δ_k is asymptotically feasible.

c) We now prove (4.16). From (4.10) and (4.18) we have for k large enough

$$\begin{aligned}\Delta_k &\geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + (d_T^k)^t H d_T^k + (d_T^k)^t(M^k - H)d_T^k + (\alpha_k - \varepsilon^2)\|d_T^k\|_2^2, \\ &\geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + (d_T^k)^t H d_T^k - 2\varepsilon^2\|d_T^k\|_2^2.\end{aligned}$$

Then using (4.12) of Lemma 4.2 with $M = H$, we obtain for all $\varepsilon > 0$

$$\Delta_k \geq \frac{\xi}{2}\|\tilde{g}(x^k)\| + (d^k)^t H d^k - 3\varepsilon^2 \|d_T^k\|_2^2 - \|H\|(1 + \|H\|/\varepsilon^2)\|d_N^k\|_2^2.$$

Hence one has from (4.17) for k large enough :

$$\Delta_k \geq (d^k)^t H d^k - 3\varepsilon^2 \|d_T^k\|_2^2. \tag{4.20}$$

Now, using the constant a_5 defined in (4.15) :

(i) if $(d^k)^t H d^k \leq a_5 \|d^k\|_2^2$, we deduce from $\dfrac{1}{2(1-\gamma)} \in (0,1)$ that (4.16) holds.

(ii) if $(d^k)^t H d^k > a_5 \|d^k\|_2^2$, we deduce from (4.20) that:

$$\Delta_k \geq \frac{1}{2(1-\gamma)}(d^k)^t H d^k + (1 - \frac{1}{2(1-\gamma)})a_5 \|d^k\|_2^2 - 3\varepsilon^2 \|d^k\|_2^2,$$

and so, choosing ε such that $3\varepsilon^2 < (1 - \dfrac{1}{2(1-\gamma)})a_5$, we get (4.16). □

Lemma 4.4 *Assume that the hypothesis of Proposition 4.1 hold. Define $\bar{x}^k := x^k - \bar{x}$. Then $\bar{x}^k = O(d^k)$.*

Proof From the optimality system of $Q_{\alpha_k}(x^k, M^k)$ we deduce that x^k satisfies the optimality system of

$$\min_x f(x) + x^t c^k \; ; \; g(x) + e^k \ll 0$$

with $c^k := M^k d^k + \alpha_k d^k$ and $e^k := g'(x^k) d^k$ and so $c^k = O(d^k)$ and $e^k = O(d^k)$.

Consider the family of perturbed problems

$$\min_x f(x) + x^t c \; ; \; g(x) + e \ll 0 \tag{$P_{c,e}$}$$

For $\bar{c} = 0$, $\bar{e} = 0$, $\bar{x}$ is a local solution of $P_{\bar{c},\bar{e}}$ satisfying the regularity hypothesis (the linearized constraints are qualified) and the strong second-order sufficient condition. It follows that for c^k, e^k close to 0, any local solution x^k of the first-order optimality system of (P_{c^k,e^k}) which is in a given neighbourhood of $\bar{x}$ is such that $\bar{x}^k = O(c^k) + O(e^k) = O(d^k)$ (see [18]). □

Proof of Proposition 4.1 : We know from Lemma 4.3 that, for k large enough, Δ_k is feasible ; so it remains to check that (4.8) holds with $\ell = 0$. Define

$$\begin{aligned} \hat{x}^{k+1} &:= x^k + d^k + v^k, \\ \check{x}^{k+1} &:= \hat{x}^{k+1} - \bar{x}, \\ a &:= \theta_r(x^k) - \theta_r(\hat{x}^{k+1}). \end{aligned}$$

We have to prove that $a \geq \gamma \Delta_k$. Indeed

$$a = L(x^k, \bar{\lambda}) - L(\hat{x}^{k+1}, \bar{\lambda}) + \bar{\lambda}^t (g(\hat{x}^{k+1}) - g(x^k)) + r(\|g(x^k)^\sharp\| - \|g(\hat{x}^{k+1})^\sharp\|). \tag{4.21}$$

Expanding $L(., \bar{\lambda})$ at $\bar{x}$ one obtains

$$L(x^k, \bar{\lambda}) - L(\hat{x}^{k+1}, \bar{\lambda}) = \frac{1}{2}(\bar{x}^k)^t H \bar{x}^k - \frac{1}{2}(\breve{x}^{k+1})^t H \breve{x}^{k+1} + o(\bar{x}^k)^2 + o(\breve{x}^{k+1})^2. \tag{4.22}$$

Moreover one has

$$\begin{aligned}(\bar{x}^k)^t H \bar{x}^k - (\breve{x}^{k+1})^t H \breve{x}^{k+1} &= (\bar{x}^k - \breve{x}^{k+1})^t H (\bar{x}^k + \breve{x}^{k+1}) \\ &= -(d^k + v^k)^t H (2\bar{x}^k + d^k + v^k).\end{aligned}$$

So using (4.4) we get :

$$(\bar{x}^k)^t H \bar{x}^k - (\breve{x}^{k+1})^t H \breve{x}^{k+1} = -2(d^k)^t H \bar{x}^k - (d^k)^t H d^k + o(d^k)^2$$

then, since (4.4) yields $\breve{x}^{k+1} = \bar{x}^k + d^k + o(d^k)$ and using Lemma 4.4, we obtain from (4.22):

$$L(x^k, \bar{\lambda}) - L(\hat{x}^{k+1}, \bar{\lambda}) = -(d^k)^t H \bar{x}^k - \frac{1}{2}(d^k)^t H d^k + o(d^k)^2.$$

Then from (4.5), (4.6) and Lemma 4.4 we get from (4.21)

$$a = -(d^k)^t H \bar{x}^k - \frac{1}{2}(d^k)^t H d^k - \bar{\lambda}^t g(x^k) + r\|g(x^k)^\sharp\| + o(d^k)^2. \tag{4.23}$$

On the other hand we have

$$\begin{aligned}\Delta_k &= r\|g(x^k)^\sharp\| - f'(x^k)d^k \\ &= r\|g(x^k)^\sharp\| - \nabla_x L(x^k, \bar{\lambda})^t d^k + \bar{\lambda}^t g'(x^k) d^k.\end{aligned}$$

So expanding $\nabla_x L(x^k, \bar{\lambda})$ at $\bar{x}$ and using Lemma 4.4 :

$$\Delta_k = r\|g(x^k)^\sharp\| - (\bar{x}^k)^t H d^k + \bar{\lambda}^t g'(x^k) d^k + o(d^k)^2.$$

Using (4.3) and the complementarity condition in (1.6), we get for any $i \in I^*$:

$$g_i(x^k) + g_i'(x^k) d^k = 0,$$

hence $-\bar{\lambda}^t g(x^k) = \bar{\lambda}^t g'(x^k) d^k$ and so :

$$\Delta_k = r\|g(x^k)^\sharp\| - (\bar{x}^k)^t H d^k - \bar{\lambda}^t g(x^k) + o(d^k)^2.$$

Plugging this in (4.23) we obtain

$$a = -\frac{1}{2}(d^k)^t H d^k + \Delta_k + o(d^k)^2$$

We want $a \geq \gamma \Delta_k$, i.e.

$$(1 - \gamma)\Delta_k \geq \frac{1}{2}(d^k)^t H d^k + o(d^k)^2$$

which reduces to relation (4.16) proved by Lemma 4.3. □

According to Section 1.4, we now present an algorithm that is globally convergent (as in Section 3) and that converges superlinearly when we assume that $\{M^k\}$ approximates in some sense the Hessian of the Lagrangian of problem (P) (using Section 4 and properties of Newton type algorithms quoted in Section 1.2). We now state the algorithm :

Algorithm 3 :

Perform the same steps as in Algorithm 2, replacing LS1 by LS2.

Theorem 4.1 *Let x^k be computed by Alg. 3. We assume that (1.7) and (1.8) hold. Then*

(i) *$\{r_k\}$ and $\{\alpha_k\}$ are bounded.*

(ii) *The set of limit points of $\{x^k\}$ is connex and to each of them is associated a Lagrange multiplier.*

(iii) *Assume that the algorithm computes the solution d^k of minimal norm of the optimality system of $Q_{\alpha_k}(x^k, M^k)$. If to some $\bar{x}$ limit-point of $\{x^k\}$ is associated a multiplier $\bar{\lambda}$ such that (4.9) and (4.10) hold, then $x^k \to \bar{x}$ and $\rho_k = 1$ for k large enough. If in addition $P^k[(\nabla_x^2 L(\bar{x}, \bar{\lambda}) - M^k)d^k] = o(d^k)$ and α_k is taken to be 0 when it it possible, i.e. when $\rho_k = 1$, then the convergence is superlinear.*

Proof : The argument for proving (i) (ii) are essentially the same as for Thm 3.1. As they are rather long we do not reproduce them in detail but rather analyse where the differences are.

Proof of (i) relies on extension of Lemmas 3.1-3.3 for Alg. 3. Lemma 3.1 is proved by checking that $\|g(x^k)^\sharp\|$ converges if $r_k \searrow \infty$, and on a first order expansion (in ρ) of $\|g(x^k + \rho d^k)^\sharp\|$. These last arguments have an immediate extensions as the paths $\rho \to x^k + \rho d^k$ and $\rho \to x^k + \rho d^k + \rho^2 v^k$ have the same first order expansion, the term v^k being uniformly bounded. Simple considerations allow an immediate extension of Lemma 3.2. For the extension of Lemma 3.3, estimate (3.5) on Δ_k is still valid, and (3.6) also holds, but with a possibly different constant a_4 (because of the additional term of v^k) and the conclusion follows. Now the same discussion of points a), b) of proof of Thm 3.1 can be used in order to check that (i) holds.

Proof of (ii) The mechanism of adaptation of $\{x^k\}$ and Lemma 3.3 imply that $\rho_{k'} > 0$ for an infinite subsequence $\{k'\}$, and we may suppose that $\{x^{k'}\} \to \hat{x}$. If $\Delta_{k'} \to 0$ it follows that $d^{k'} \to 0$ hence $\hat{x}$ is a stationary point of (P). If not, assuming $d^{k'} \to \hat{d} \neq 0$ and $v^{k'} \to \hat{v}$ (note that $v^{k'}$ is bounded by (4.4) hence has limit-points) expanding $\rho \to \theta_r(\hat{x} + \rho\hat{d} + \rho^2\hat{v})$ as in the proof of Thm 2.1. we deduce that $\rho_{k'}$ cannot converge to 0, hence $\theta_r(x^{k'}) \to \infty$, which is impossible. Henceforth $\hat{d} = 0$ and point (ii) follows. Using (4.9), (4.10), and applying the sensitivity result of Robinson [18] to $\bar{d} = 0$ solution of $Q(\bar{x}, \nabla_x^2 L(\bar{x}, \lambda))$ we deduce that $d^k \to 0$ for the considered subsequence.

Proof of (iii) That $\rho_{k'} = 1$ asymptotically for the subsequence $\{x^{k'}\} \to \bar{x}$ is then a consequence of Prop 4.1. Indeed $\mu^{k'} \to \bar{\lambda}$ as $d^{k'} \to 0$ and (M^k, α_k) are bounded. If $\rho_{k'} < 1$ for a subsequence then (for k' large enough) $r_{k'+1} > \|\bar{\lambda}\|_*$, hence $r > \|\bar{\lambda}\|_*$ and the hypotheses of Prop 4.1 are satisfied : it follows that $\rho_{k'} = 1$ for k' large enough.

Now by (4.9), $\bar{x}$ is an isolated stationary point (see Robinson [18]), and by point (ii) is an isolated limit-point of $\{x^k\}$. As the set of limit points of $\{x^k\}$ is connex it follows that all the sequence converges to $\bar{x}$.

If in addition $P^k[(\nabla_x^2 L(\bar{x},\bar{\lambda}) - M^k)d^k] = o(d^k)$, then by Thm 1.2, $x^k + d^k - \bar{x} = o(x^k - \bar{x})$. As $v^k = O(d^k)^2 = o(x^k - \bar{x})$ we get $x^{k+1} - \bar{x} = o(x^k - \bar{x})$, as desired. □

We now formulate an algorithm that, assuming that the second derivatives of f and g are known, is an extension of Newton's method in the sense that, when x^k is close to some $\bar{x}$ satisfying (4.9), it computes d^k using $M^k = \nabla_x^2 L(x^k, \mu^{k-1})$ where μ^{k-1} is the multiplier associated to d^{k-1}, and $x^k \to \bar{x}$ with a quadratic rate. The rule is as follows :

$$\left\{ \begin{array}{l} \text{choose } M^{k+1} = \nabla_x^2 L(x^{k+1}, \lambda^{k+1}) \text{ with} \\ \\ \lambda^{k+1} := \left\{ \begin{array}{ll} \mu^k & \text{if } \alpha_k \|d^k\| + \|M^k d^k\| \le 1, \\ \\ \mu^k/(1 + \alpha_k\|d^k\| + \|M^k d^k\|) & \text{if not.} \end{array} \right. \end{array} \right. \tag{4.24}$$

Theorem 4.2

a) Let $\{x^k\}$ be computed by Algorithm 3 with $\{M^k\}$ computed by (4.24). We assume that $\{x^k\}$, $\{d^k\}$ are bounded, that (1.8) holds and that $\alpha_{k+1} = 0$ if $\rho_k = 1$. Then points (i) (ii) of Thm 4.1 still hold.

b) In addition, if $\bar{x}$ satisfying (4.9) is limit-point of x^k and d^k is the solution of minimal norm of the optimality system of $Q_{\alpha_k}(x^k, M^k)$, then all the sequence $\{x^k\}$ converges to $\bar{x}$ with a quadratic rate.

Proof a) In order to get point (i) (ii) of Thm 4.1 we just have to check that $\{M^k\}$ is bounded ; Indeed λ^{k+1} is bounded by (1.8) and (4.24) hence so is $\{M^k\}$.

Now as $d^k \to 0$ and (M^k, α_k) are bounded, it follows that $\mu^k \to \bar{\lambda}$ and $\lambda^{k+1} = \mu^k$ by (4.24), hence $M^k \to \nabla_x^2 L(\bar{x}, \bar{\lambda})$ and point (iii) of Thm 4.1 implies that $\rho_k = 1$ since (4.10) obviously holds which implies the convergence of all the sequence to $\bar{x}$ at a quadratic rate by Theorem 1.2. □

References

[1] Armijo L (1966) : *Minimization of function having Lipschitz continuous first partial derivatives.* Pacific J. Math. 16, 1-3.

[2] Ben-Tal A (1980) : *Second-order and related extremality conditions in nonlinear programming.* J. Optimiz. Theory Appl. 21, 143-165.

[3] Bertsekas DP (1982) : *Constrained optimization and Lagrange multipliers methods.* Academic Press, New York.

[4] Boggs JW, Tolle PW, Wang P(1982) : *On the local convergence of quasi-Newton methods for constrained optimization.* SIAM J. Control Optimiz. 20, 161-171.

[5] Bonnans JF (1989) : *Local study of Newton type algorithms for constrained problems.* Proc. 5th Franco-German conf. on Optimization, S. Dolecki ed., Lectures notes in Math. 1405, Springer-Verlag, 13-24.

[6] Bonnans JF (1989) : *Asymptotic admissibility of the unit stepsize in exact penalty methods.* SIAM J. Control Optimization 27, 631-641.

[7] Bonnans JF : *Local analysis of Newton type methods for variational inequalities and nonlinear programming.* To appear.

[8] Bonnans JF, Panier ER, Tits AL, Zhou JL (1992) : *Avoiding the Maratos effect by means of a nonmonotone line search II : Inequality constrained problems - feasible iterates.* SIAM J. Numerical Analysis, 29, 1187-1202.

[9] Chamberlain RM, Lemaréchal C, Pedersen HC, Powell MJD (1982) : *The watchdog technique for forcing convergence in algorithm for constrained optimization.* Math. Programming Stud. 16, 1-17.

[10] Fletcher R (1987) : *Practical methods of optimization.* (2nd edition) Wiley, Chichester.

[11] Gabay D (1982) : *Reduced quasi-Newton methods with feasibility improvement for nonlinearly constrained optimization.* Math. Programming Stud. 16, 18-44.

[12] Han P (1977) : *A globally convergent method for nonlinear programming.* J. Optim. Theory Appl. 22, 297-309.

[13] Han P, Mangasarian OL (1979) : *Exact penalty functions in nonlinear programming.* Math. Programming, 17, 251-269.

[14] Maratos N (1978) : *Exact penalty function algorithms for finite dimensional and control optimization problems.* Ph. D. thesis, University of London, London, U.K.

[15] Mayne DQ, Polak E (1982) : *A superlinearly convergent algorithm for constrained optimization problems.* Math. Programming Stud. 16, 45-61.

[16] Panier ER, Tits AL : *Avoiding the Maratos effect by means of a nonmonotone line search I : General constrained problems.* SIAM J. Numerical Analysis, 28(1991), 1183-1195.

[17] Pschenichny BN, Danilin YM (1975) : *Numerical methods in extremal problems,* MIR, Moscow (Engl. Trans. 1978).

[18] Robinson SM (1982) : *Generalized equations and their applications, Part II : application to nonlinear programming* Math. Programming Study 19, 200-221.

[19] Wilson RB : *A simplicial algorithms for concave programming*, Ph. D. dissertation, Harvard University, 1963.

[20] Yuan Y : *An example of only linear convergence of trust region algorithms for non-smooth optimization,* IMA J. Numer. Anal. 4(1984), 327-335.

ISSN 0249-6399